UNIVERSITÀ
DEGLI STUDI
DE L'AQUILA

Diana Project

Corso di Bio-Informatica

Andrea Serafini, Lorenzo Andreoli, Paolo Tramontozzi, Giuseppe Gasbarro

AA 2020-2021

SOMMARIO

1 GLOSSARIO

Cromosoma	il cromosoma è la struttura con cui, durante il processo riproduttivo della cellula, ciascuna unità funzionale di DNA, dopo essersi duplicata, si compatta associata a specifiche proteine e viene trasmessa alle cellule figlie.
DNA	L'acido desossiribonucleico o deossiribonucleico (DNA) è un acido nucleico che contiene le informazioni genetiche necessarie alla biosintesi di RNA e proteine, molecole indispensabili per lo sviluppo ed il corretto funzionamento della maggior parte degli organismi viventi.
Eteroduplex	L'*eteroduplex* è un termine della genetica che indica una struttura ibrida di una molecola di acido nucleico e si riferisce all'area formata da due filamenti provenienti da molecole diverse. Un esempio di *eteroduplex* è quello che si crea al termine della trasformazione batterica nella quale il filamento esogeno si inserisce nel cromosoma batterico della cellula accettrice che ne crea il complementare.
Geni	I geni corrispondono a porzioni di genoma localizzate in precise posizioni all'interno della sequenza di DNA (o più raramente RNA in certi virus) e contengono le informazioni necessarie per la produzione di una proteina o di RNA (senza la produzione di proteine).
Genoma	Il genoma è l'insieme di tutte le informazioni genetiche depositate nella sequenza del DNA contenuto nel nucleo delle cellule sotto forma di cromosomi. Ogni cromosoma è costituito da un lungo filamento di DNA organizzato in una complessa struttura tridimensionale.
microRNA	I microRNA *(miRNA)* sono piccole molecole endogene di RNA non codificante a singolo filamento riscontrate nel trascrittoma di piante, animali ed alcuni virus.
Predizione di struttura proteica	Per predizione di struttura proteica *(protein structure prediction)* s'intende la predizione della struttura tridimensionale d'una proteina, a partire dalla sua sequenza aminoacidica, ossia la predizione della sua struttura secondaria, ternaria, quaternaria, partendo dalla sua struttura primaria.
Proteina	Le proteine o protidi sono grandi biomolecole o macromolecole, costituite da una o più catene di aminoacidi. Le proteine svolgono una vasta gamma di funzioni all'interno degli organismi viventi, compresa la catalisi delle reazioni metaboliche, la replicazione del DNA, la risposta agli stimoli e il trasporto di molecole da un luogo ad un altro. Le proteine differiscono l'uno dall'altra soprattutto nella loro sequenza di amminoacidi, la quale è dettata dalla sequenza nucleotidica conservata nei geni e che di solito si traduce in un ripiegamento proteico in una struttura tridimensionale specifica che determina la sua attività.
RefSeq	NCBI's Reference Sequence (RefSeq) database è una collezione non ridondante che rappresenta le molecole che compongonoDNA,RNA e proteine.
RNA	L'RNA, o acido ribonucleico, è una molecola polimerica implicata in vari ruoli biologici di codifica, decodifica, regolazione e l'espressione dei geni. L'RNA e il DNA sono acidi nucleici, e, insieme a proteine e carboidrati, costituiscono le tre principali macromolecole essenziali per tutte le forme di vita conosciute.
UTR	Con il termine Untranslated region (solitamente abbreviato con UTR) si indicano delle sequenze localizzate alle estremità 5' e 3' di un RNA messaggero (acido ribonucleico codificante per una proteina) che non vengono tradotte, sono denominate rispettivamente 5' UTR e 3' UTR.

2 INTRODUZIONE

Il Diana Project è un applicativo software focalizzato sullo studio del genoma del topo domestico (Mus Musculus), in grado di combinare le informazioni contenute in diverse banche, e tracciare le relazioni tra miRNA ed i Geni target.

L' obiettivo per l'AA 2020-2021 è di aggiornare e migliorare il vecchio applicativo del DIANA Project dell'AA 2015/2016 (A. di Marco, 2016) (Tucci Michele, 2016)

Esso, come già accennato, permette di tracciare le relazioni tra microRNA e Geni della specie *Mus Musculus (topo domestico).*

I dati relativi al sequenziamento del genoma del topo sono pubblici e resi disponibili da diverse banche dati; per questo elaborato sono state utilizzate PicTar, RNA22, TargetScan e miRtarBase.

Le migliorie principali del progetto per l'AA 2020-2021 riguardano: definire un'architettura di sistema che sia facilmente mantenibile per aggiornamenti futuri, verificare la disponibilità dei dati utilizzati, aggiornando il database, e fornire all'utente finale un'interfaccia che permetta di interrogare facilmente il database.

3 ARCHITETTURA

La nuova architettura è costituita da 3 layer principali in linea con il pattern MVC. Il modello è basato su una rappresentazione dei dati a grafo, ed è implementato con Neo4J. I dati vengono generati attraverso diversi script Python che si occupano del parsing dei dati delle diverse banche dati utilizzate e che effettuano anche le query per la costruzione del DB utilizzato poi dall'applicazione web Diana. Questo processo viene effettuato una sola volta durante la costruzione del DB a grafo, mentre successivamente l'applicazione utilizzerà solo il DB Neo4J. Il funzionamento dell'applicazione web Diana è il classico dei framework basati sul pattern MVC, dove l'utente, tramite una form del View Layer, invia una richiesta che viene gestita dal Controller Layer di Flask, quindi recupera i dati nel Model Layer e li restituisce all'utente su una nuova View.

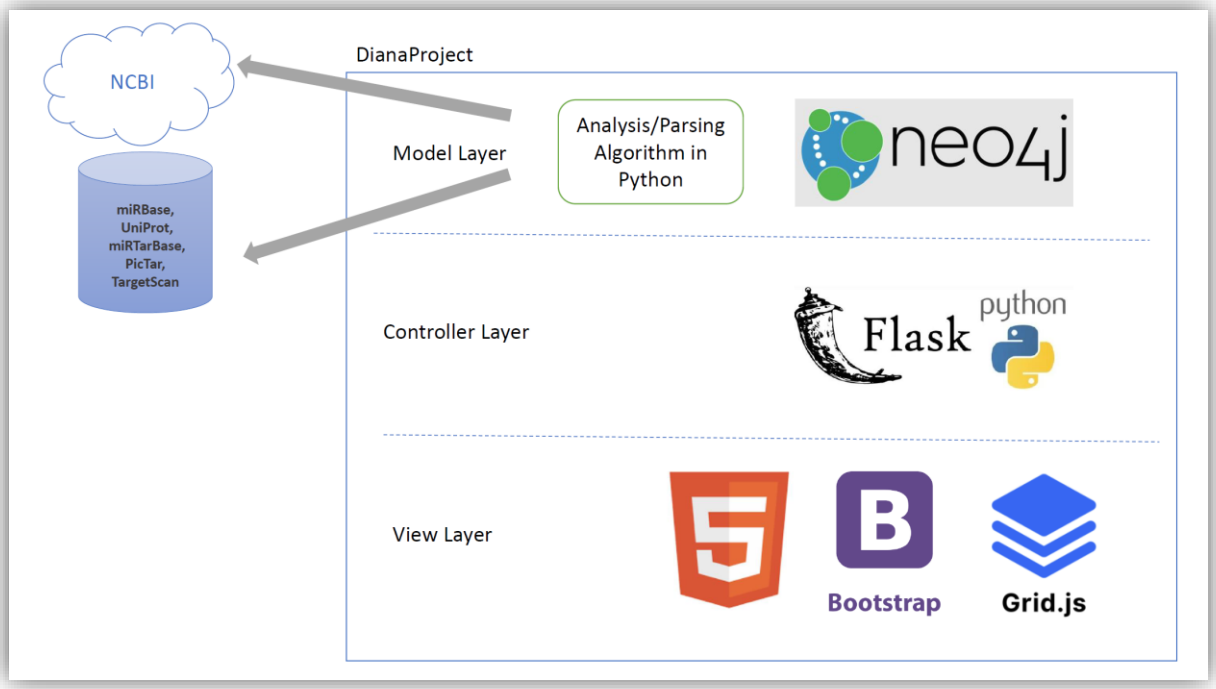


Figura 1 nuova architettura software Diana project

La fase di costruzione del database è quella più onerosa in quanto consiste nel parsing di file con anche 60 milioni di entry. Questa fase preliminare viene implementata con diversi script

scritti in Python. I vecchi file sono stati modificati e resi compatibili con le nuove versioni di Python e Neo4J.

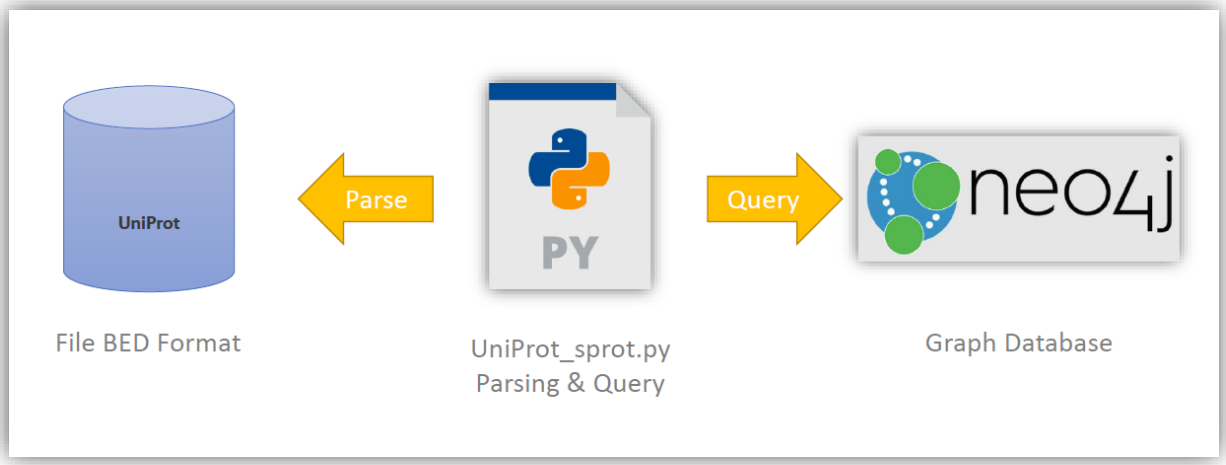


Figura 2 nella figura precedente il dettaglio del della costruzione del database a partire dai file delle banche dati.

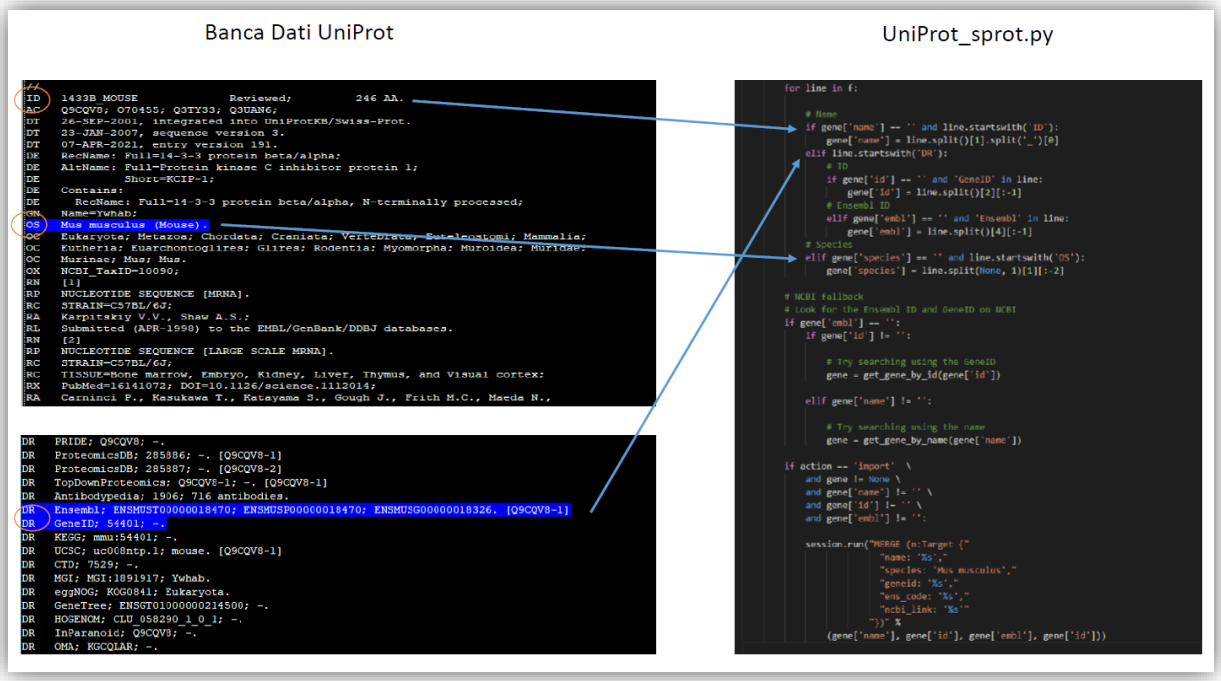


Figura 3 in figura è riportato un esempio di file python che fa il parse della corrispondente banca dati, argomento approfondito nel capitolo Banche Dati & Data Mapping.

Tali script, dal punto di vista logico, implementano una mappatura tra i campi presenti nei file delle banche dati ed il nuovo database, argomento che verrà trattato in dettaglio nel prossimo capitolo.

4 DATABASE NEO4J

Neo4j è un database a grafo caratterizzato dalle alte performance e con tutte le funzionalità che ci si aspetterebbe da una base di dati robusta (es. un linguaggio query, ACID transactions). In questo modo il programmatore si interfaccia con una struttura composta da nodi e relazioni, piuttosto che tabelle.

Inoltre, per molte applicazioni, Neo4j offre prestazioni notevoli, beneficiandone l'utilizzo, in contrasto con quelle dei classici database relazionali. In particolare, per le esigenze di questo progetto, la grande mole di dati e la necessità dell'utilizzo di "join" in caso di database relazionale (che avrebbero impattato negativamente sulle performance) hanno fatto ricadere la scelta su Neo4J.

Infatti, da come possiamo notare dalla seguente citazione e dai benchmark sottostanti, un database a grafi fornisce quasi sempre una struttura adatta alle richieste delle applicazioni bioinformatiche.

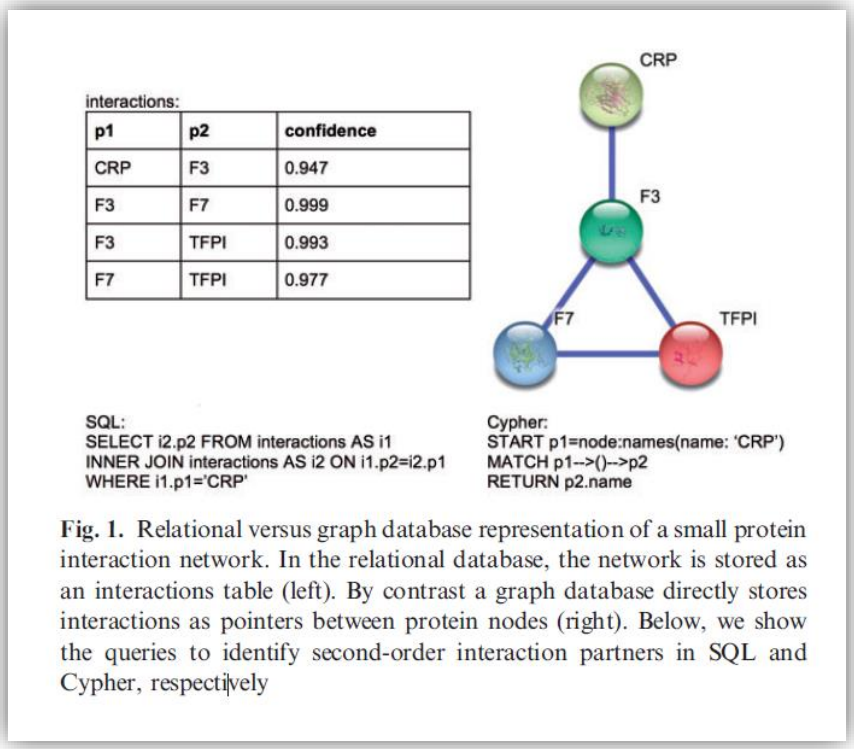


Fig. 1. Relational versus graph database representation of a small protein interaction network. In the relational database, the network is stored as an interactions table (left). By contrast a graph database directly stores interactions as pointers between protein nodes (right). Below, we show the queries to identify second-order interaction partners in SQL and Cypher, respectively

Table 1. Query benchmark of a relational and a graph database

	Neighbor network	Best-scoring path	Shortest path
PostgreSQL	206.31 s	1147.74 s	976.22 s
Neo4j	5.68 s[a]	1.17 s	0.40 s
Speedup	36×	981×	2441×

Figura 4 Le figure precedenti mostrano un confronto tra i database SQL e quelli a grafo. (Are graph databases ready for bioinformatics?, 2013)

> In summary, graph databases themselves are ready for bioinformatics and can offer great speedups over relational databases on selected problems. The fact that a certain dataset is a graph, however, does not necessarily imply that a graph database is the best choice; it depends on the exact types of queries that need to be performed. Graph queries formulated in terms of paths can be concise and intuitive compared with equivalent SQL queries complicated by joins.
>
> (Are graph databases ready for bioinformatics?, 2013)

4.1 GENERAZIONE DEI NODI

microRNA (miRBase)	**Target** (UniProt)
• Id • Name • Synonyms • Accession • Species • Mirbase_link	• Id • Name • Geneid • Ens_code • Species • Ncbi_link

Nel caso dei **microRNA** si hanno:

- **Id:** contiene il valore univoco generato da *Neo4J* durante la creazione del nodo;
- **Name:** conserva il nome del microRNA;
- **Synonyms:** contiene i sinonimi dei microRNA presenti in letteratura ;
- **Accession:** contiene l'identificatore stabile univoco di *miRBase*;
- **Species:** identifica la specie (ad esempio, *Mus Musculus*);
- **Mirbase_link:** ospita il codice *accession* usato per costruire il link a *miRBase*.

Rispettivamente **Target** contiene:

- **Id:** contiene il codice univoco prodotto da *Neo4J*;
- **Name:** conserva il nome del gene;
- **Geneid:** contiene l'identificativo del gene usato da NCBI;
- **Ens_Code:** contiene il codice *Ensembl*;
- **Species:** identifica la specie;
- **Ncbi_link:** contiene il *Geneid* usato per costruire il link a NCBI.

4.2 RELAZIONI TRA NODI

Le relazioni permettono di rilevare il legame tra un gene microRNA e un nodo target.

La misurazione avviene attraverso lo score. Ad ogni relazione è associata un'etichetta (label), che identifica il nome del database di appartenenza e uno o più valori specifici della relazione stessa.

Relation
• ID • Name • Score • Source_MicroRNA • Source_target

La struttura Relation presente in Neo4J contiene i seguenti campi:

- **ID:** generato da *Neo4J*;
- **Name:** indica il nome del database al quale appartiene;
- **Score:** identifica il punteggio ottenuto dalla similitudine tra due sequenze;
- **Source_MicroRNA:** contiene il nome invariato del microRNA presente nel database di origine;
- **Source_target:** conserva il target del database sorgente.

4.3 NODI ADDIZIONALI

La base di dati prevede inoltre altri tre tipi di nodi: External_link, DB_info e Relation_general_info.

External_link	**DB_info**	**Relation_general_info**
• Id • Name • Url	• Id • Name • Link	• Id • Name • source_db_link • min_value • max_value • cut_off

Le tabelle mostrate sopra contengono rispettivamente:

- gli URL utilizzati per la creazione di link esterni (*NCBI*, *MirBase*, etc.) ;
- informazioni sul database analizzato con il relativo link;
- informazioni generali sulle relazioni con i corrispondenti valori di oscillazione conservati nei campi *min_value* e *max_value*. L'attributo *cut_off* indica, invece, il valore minimo dello score assegnato alla predizione per essere ritenuto significativo.

5 BANCHE DATI & DATA MAP

Le banche dati Bioinformatiche contengono informazioni riguardanti il sequenziamento di genomi di esseri viventi di diverse specie (uomo, topo domestico, etc). In queste banche dati sono stati collezionati moltissimi dati nel corso degli anni, in alcuni casi sono nate più di 20 anni fa (Kegg 1995), questo comporta un problema di standardizzazione della rappresentazione dei dati, uno degli obiettivi del progetto Diana è uniformare la rappresentazione di dati e costruire un unico database gestito tramite Neo4J.

Tra le più affidabili basi di dati bioinformatiche possiamo si possono citare, PicTar, TargetScan, Miranada, Mirtarbase, questa in particolare sono state citate da molti articoli scientifici (A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions, 2008) *e* (miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database di Hsi-Yuan Huang Et Al., 2020).

Tali banche dati sono state scaricate ed è stato effettuato un nuovo parse dei dati, al fine di integrare gli aggiornamenti dei dati avvenuti negli ultimi anni.

5.1 PICTAR

PicTar è un algoritmo per l'identificazione di target microRNA, dove dato un miRNA conservato in allineamenti multipli di genomi di specie diverse e un set di sequenze di 3' UTR ortologhe, esegue i seguenti passi:

1. Usa il seed di 7 nt per trovare tutti match perfetti imperfetti nelle UTR.
2. Predice l'energia libera ottimale dell'ibrido.
3. Calcola la probabilità che la sequenza del target sia un sito di legame.
4. Valuta favorevolmente la presenza di più siti legame sulla stessa UTR.

Tale algoritmo viene utilizzato per calcolare il valore del campo *score* all'interno della base di dati PicTar (Fig.1). Da questa banca dati è possibile ottenere vari dataset a seconda del livello di conservazione del Mus Musculus in esame:

- conservazione tra sette vertebrati: topo, ratto, coniglio, umano, scimpanzé, macaco e cane (picTarMiRNADog_mm7.bed)
- conservazione tra tredici vertebrati: topo, ratto, coniglio, umano, scimpanzé, macaco, cane, mucca, armadillo, elefante, tenrecidi, opossum e pollo (p icTarMiRNAChicken_mm7.bed).

Non avendo delle API che consentissero l'interrogazione diretta in modo agile, ma solamente una form HTML per interrogare i datasets, le informazioni sono state scaricate direttamente mediante **UCSC Genome Browser.** In questo modo, vengono scaricati dei file nel formato *BED* che permette una classificazione dei dati migliore, meglio strutturata e facilmente processabile come Tab Separated Values.

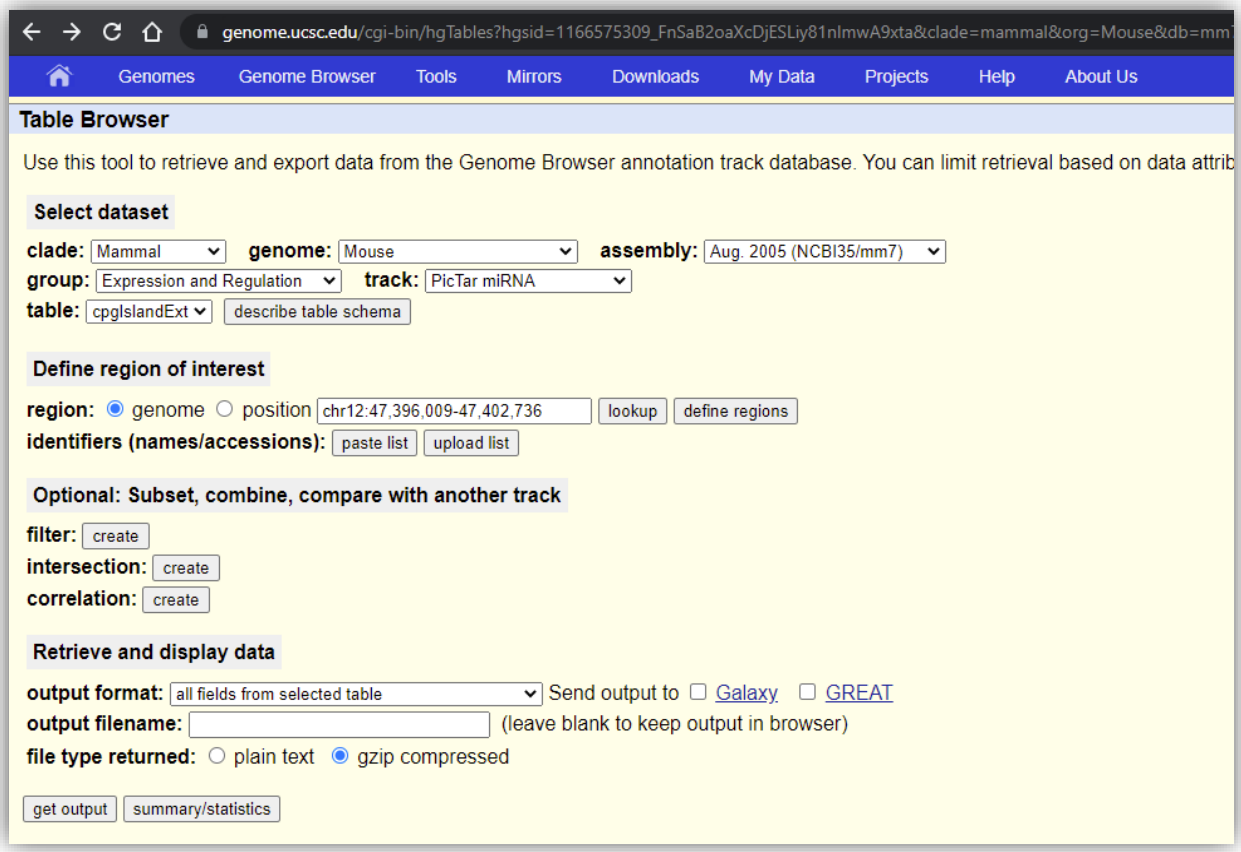


Figura 5 Interfaccia per il download del database PicTar

Il File PicTar

#bin	chrom	chromStart	chromEnd	name	score	strand	thickStart	thickEnd	
609	chr1	3211381	3211388	NM_001011874:mmu-miR-155	20	-	3211381	3211388	0
609	chr1	3211394	3211401	NM_001011874:mmu-miR-29b	38	-	3211394	3211401	0
609	chr1	3211395	3211402	NM_001011874:mmu-miR-29c	29	-	3211395	3211402	0
609	chr1	3211395	3211402	NM_001011874:mmu-miR-29a	26	-	3211395	3211402	0
609	chr1	3211451	3211458	NM_001011874:mmu-miR-202	29	-	3211451	3211458	0
609	chr1	3211470	3211477	NM_001011874:mmu-miR-138	115	-	3211470	3211477	0
609	chr1	3211515	3211522	NM_001011874:mmu-miR-138	115	-	3211515	3211522	0

Figura 6 PicTar file format

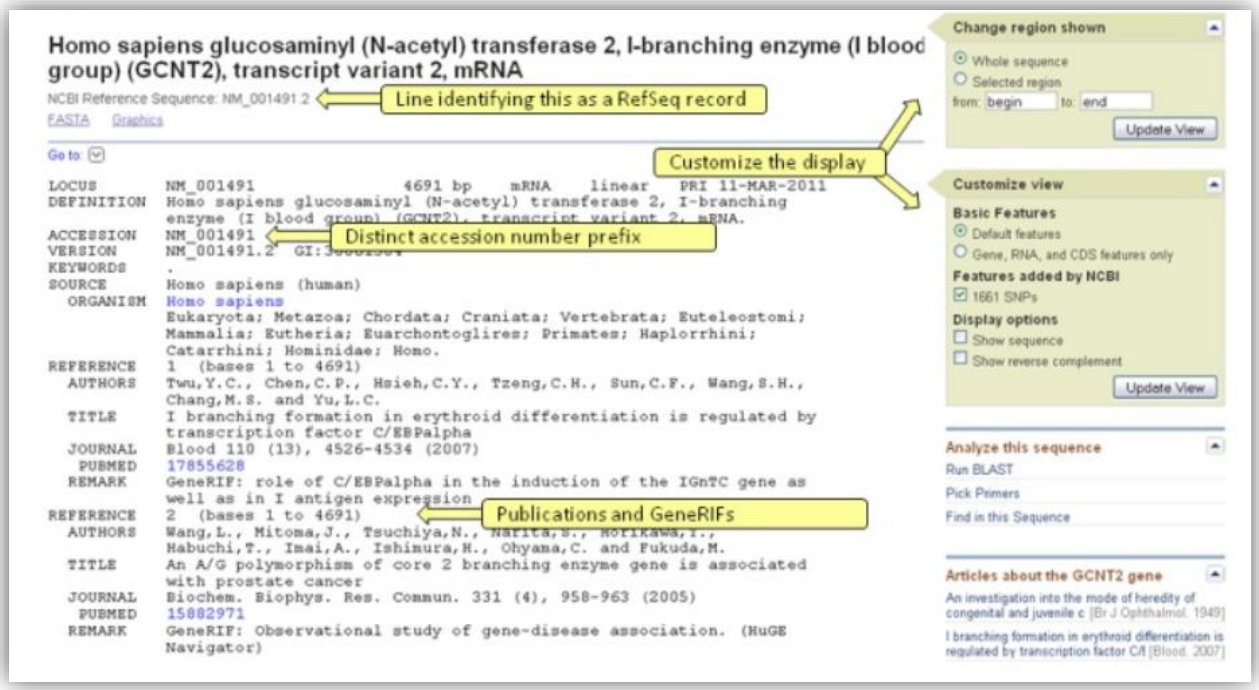


Figura 7 nella precedente figuraè riportato un riferimento al RefSeq di NCBI identificativo non ridondante di sequenze di DNA, RNA o proteine.

I campi di interesse per il database Pictar sono riportati nella seguente tabella.

name	La proprietà 'name' delle relazioni 'PicTar' assume i valori 'PicTar7' e 'PicTar13' ad indicare rispettivamente il livello di conservazione in sette e in tredici specie.
microRNA	Ottenuto tramite la stringa che segue il carattere ':' nella quinta colonna. Esso viene utilizzato per ottenere il codice *accession* che permette di ottenere un match preciso e indipendente dalle piccole variazioni dei nomi dei microRNA e quindi utilizzato per cercare e ricavare il nodo microRNA in Neo4j.
target	La stringa che precede ':', sempre nella quinta colonna, si riferisce al target. Dato che PicTar si riferisce a tale target attraverso il codice RefSeq del nucleotide corrispondente, esso viene usato per trovare il GeneID interrogando NCBI. Tale GeneID viene ottenuto per trovare il nodo target in Neo4j. In caso non venga trovato ne viene creato uno utilizzando le informazioni ricavate dall'interrogazione precedente verso NCBI.
score	Quinta colonna.

Dettagli Banca Dati PicTar

Database	Descrizione	Download
PicTar	picTarMiRNADog_mm7.bed picTarMiRNAChicken_mm7.bed - - BED format	https://genome.ucsc.edu/cgi-bin/hgTables clade: Mammal genome: Mouse assembly: Aug. 2005 (NCBI35/mm7) group: Expression and Regulation track: PicTar miRNA table: PicTar 7 Species (picTarMiRNADog) and PicTar 13 Species (picTarMiRNAChicken)

5.2 UNIPROT

La banca dati uniProt contiene sequenze proteiche di alta qualità, questa è suddivisa in due sezioni, UniProtKB/Swiss-Prot costruita e revisionata manualmente e UniProtKB/TrEmbl costruita su voci revisionate e annotate automaticamente via software. UniProt è considerata una delle migliori banche dati fin dai primi anni 2000.

> *THE UNIPROT ARCHIVE (UNIPARC) The UniProt Archive (UniParc) is the most comprehensive publicly accessible non-redundant protein sequence collection available. It contains publicly available protein sequences from many different sources, including Swiss-Prot, TrEMBL, PIR-PSD, EMBL (3), Ensembl (4), IPI (http://www.ebi.ac.uk/ IPI), PDB (5), RefSeq (6), FlyBase (7), WormBase (8), and European, American and Japanese patent offices. While a protein sequence may exist in multiple databases and more than once in a given database, UniParc stores each unique sequence only once and assigns a unique UniParc identifier.*
>
> *To provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information, the Swiss-Prot, TrEMBL and PIR protein database activities have united to form the Universal Protein Knowledgebase (UniProt) consortium. Our mission is to provide a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and query interfaces.*
>
> *(UniProt: the Universal Protein knowledgebase, 2004)*

Il database UniProt fornisce un unico grande file (uniprot_sprot.dat) di notevoli dimensioni (circa 2.8GB) che non permettono un processamento diretto abbastanza veloce (ad esempio per isolare una specie di interesse). È stato quindi diviso in numerosi file di piccole dimensioni ognuno contenente un blocco di dati (un gene). Ogni identificativo è composto dal nome del gene seguito dal carattere '_', seguito dal nome della specie (ad esempio, 'MOUSE' per 'Mus Musculus').

Il File Uniprot

```
//
ID   1433B_MOUSE             Reviewed;         246 AA.
AC   Q9CQV8; O70455; Q3TY33; Q3UAN6;
DT   26-SEP-2001, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 3.
DT   07-APR-2021, entry version 191.
DE   RecName: Full=14-3-3 protein beta/alpha;
DE   AltName: Full=Protein kinase C inhibitor protein 1;
DE            Short=KCIP-1;
DE   Contains:
DE     RecName: Full=14-3-3 protein beta/alpha, N-terminally processed;
GN   Name=Ywhab;
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae;
```

```
//
ID   1433B_MOUSE             Reviewed;         246 AA.
AC   Q9CQV8; O70455; Q3TY33; Q3UAN6;
DT   26-SEP-2001, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 3.
DT   07-APR-2021, entry version 191.
DE   RecName: Full=14-3-3 protein beta/alpha;
DE   AltName: Full=Protein kinase C inhibitor protein 1;
DE            Short=KCIP-1;
DE   Contains:
DE     RecName: Full=14-3-3 protein beta/alpha, N-terminally processed;
GN   Name=Ywhab;
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae;
OC   Murinae; Mus; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   STRAIN=C57BL/6J;
RA   Karpitskiy V.V., Shaw A.S.;
RL   Submitted (APR-1998) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RC   STRAIN=C57BL/6J;
RC   TISSUE=Bone marrow, Embryo, Kidney, Liver, Thymus, and Visual cortex;
RX   PubMed=16141072; DOI=10.1126/science.1112014;
RA   Carninci P., Kasukawa T., Katayama S., Gough J., Frith M.C., Maeda N.,
```

```
DR   PRIDE; Q9CQV8; -.
DR   ProteomicsDB; 285886; -. [Q9CQV8-1]
DR   ProteomicsDB; 285887; -. [Q9CQV8-2]
DR   TopDownProteomics; Q9CQV8-1; -. [Q9CQV8-1]
DR   Antibodypedia; 1906; 716 antibodies.
DR   Ensembl; ENSMUST00000018470; ENSMUSP00000018470; ENSMUSG00000018326. [Q9CQV8-1]
DR   GeneID; 54401; -.
DR   KEGG; mmu:54401; -.
DR   UCSC; uc008ntp.1; mouse. [Q9CQV8-1]
DR   CTD; 7529; -.
DR   MGI; MGI:1891917; Ywhab.
DR   eggNOG; KOG0841; Eukaryota.
DR   GeneTree; ENSGT01000000214500; -.
DR   HOGENOM; CLU_058290_1_0_1; -.
DR   InParanoid; Q9CQV8; -.
DR   OMA; KGCQLAR; -.
```

Figura 8 Uniprot File Format

I campi di interesse per il database Uniprot sono riportati nella seguente tabella.

name	riga che inizia con 'ID', il nome del gene è la prima stringa dopo il primo TAB, solo la parte prima del '_';
species	riga che inizia con 'OS', le prime due stringhe separate da spazio (esclude la parte tra parentesi);
ens_code	riga contenente 'Ensembl', quinta colonna escludendo il '.' finale;
geneid	riga contenente 'GeneID', terza colonna escludendo il ';' finale;
ncbi_link	viene semplicemente riportato il GeneID utile a costruire il link a NCBI.

Dettagli Banca Dati UniProt

Database	Descrizione	Download
UniProt (Swiss-Prot)	UniProtKB - Reviewed (Swiss-Prot) - text format	ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz

5.3 RNA22

RNA22 è un algoritmo *pattern-based* in grado di trovare siti di legami di microRNA nella sequenza d'interesse e successivamente di identificare il target. Per l'identificazione utilizza una particolare metodologia per la previsione da microRNA a mRNA *eteroduplex*. Infatti, per la formazione, non vengono utilizzate *eteroduplex* validate sperimentalmente, ma sequenze di miRNA maturi presenti nelle banche dati pubbliche. Una volta individuato un microRNA target, tramite la ricerca di pattern, quest'ultimo può essere identificato grazie a uno dei vari algoritmi disponibili in grado di prevedere le trascrizioni da microRNA a mRNA *eteroduplex*.

RNA22 contiene le previsioni per tutte le trascrizioni di codifica delle proteine delle seguenti specie:

- Homo Sapiens;
- Mus Musculus;
- Drosophila Melanogaster;
- Caenorhabditis Elegans.

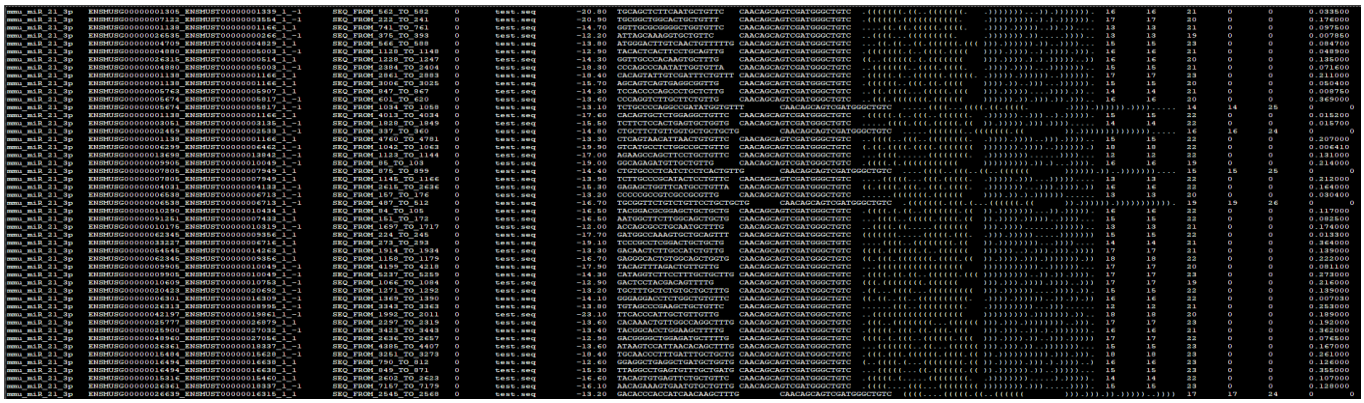
Vengono fornite tre metodologie differenti per ottenere le informazioni che esso contiene:

- predizione precalcolata;
- permette all'utilizzatore di immettere le proprie sequenze in modo personalizzato;
- permette di ottenere l'intero dataset.

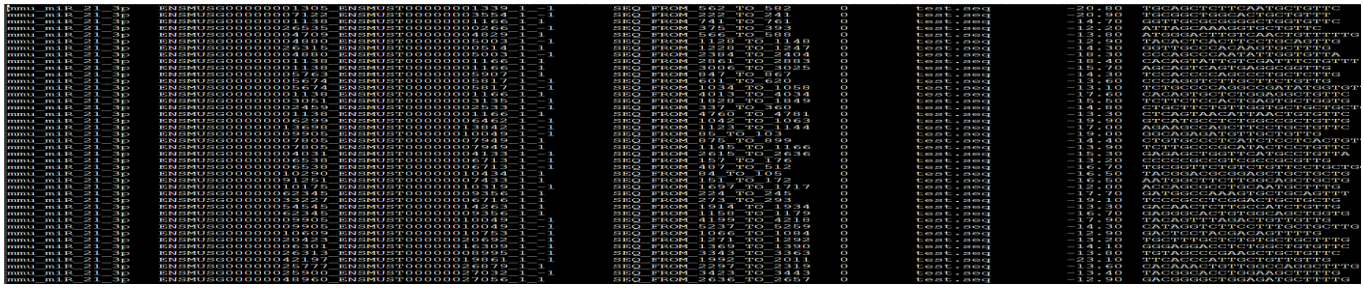
Nello studio condotto si è scelto di optare per la terza opzione, ovvero procedendo con il download dell'intero dataset della specie Mus Musculus.

Le predizioni di RNA22 vengono fornite separate per specie. Nel caso del Mus Musculus, queste corrispondono ad un dataset composto da 1157 file in formato Tab Separated Values, uno per ogni microRNA, con i relativi target per un totale di 64.528.776 record.

Il File Ensembl di RNA22



Parte sinistra



Parte destra

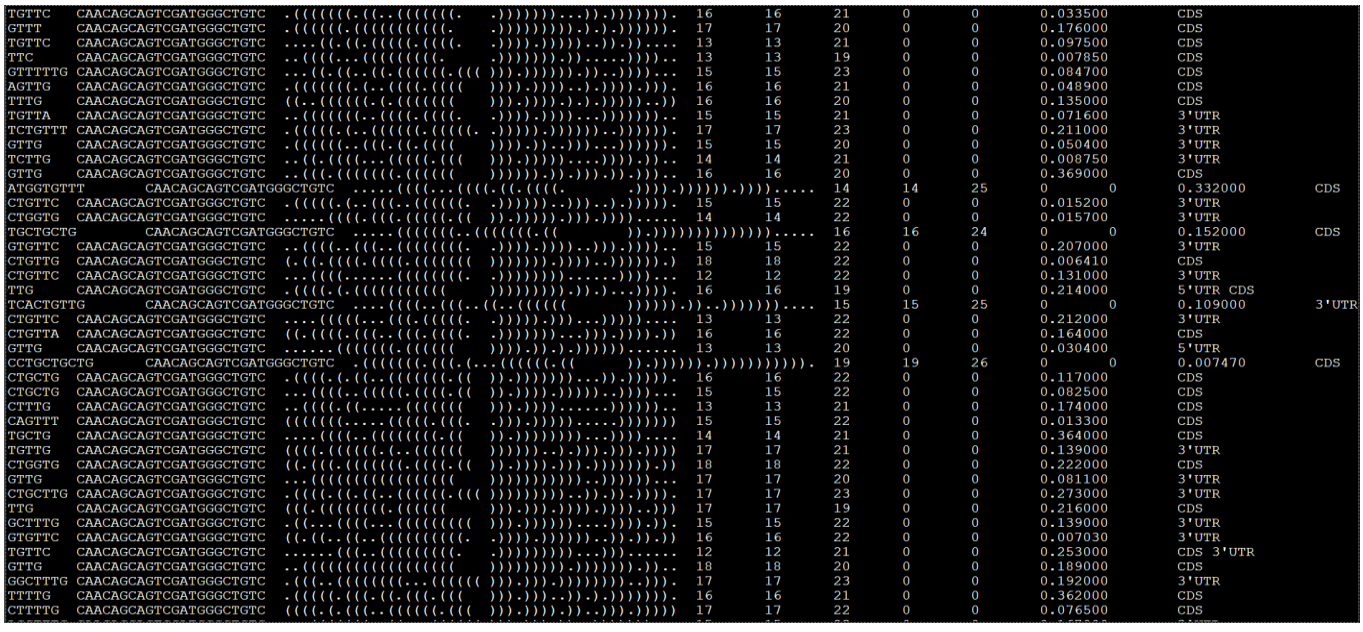


Figura 9 Ensembl File

I campi di interesse per il database RNA22 sono riportati nella seguente tabella.

name	nome fisso della relazione indicante la versione;
microRNA	prima colonna. Nel nome del microRNA fornito da RNA22 vengono rimpiazzati i '_' con '-' al fine di trovare corrispondenze (case-insensitive) con i nodi importati da miRBase. Quando questo non è possibile si prova ad ottenere un codice accession cercando il nome nel file 'aliases.txt' contenente una lista di alias fornita da miRBase. Fallito anche questo tentativo, viene creato un nuovo nodo microRNA;
Target	seconda colonna, stringa che precede '_'. Si prova a cercare un nodo Target corrispondente utilizzando il codice Ensembl. Quando il nodo non è presente se ne crea uno nuovo cercando le informazioni relative (nome, GeneID, specie) interrogando nell'ordine NCBI, UniProt e ensembl.org;
score	sesta colonna;
source_microrna	contiene il nome del microRNA così come fornito da RNA22;
source_target	contiene il codice Ensembl.

Dettagli Banca Dati RNA22

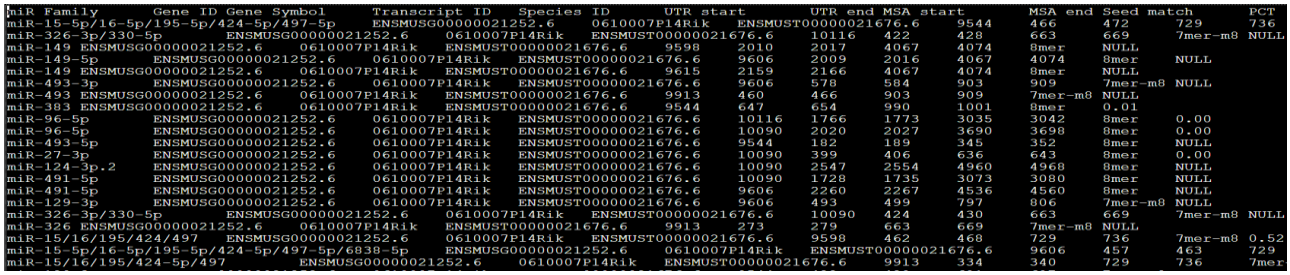
Database	Descrizione	Download
RNA22	MusMusculus,mRNA,ENSEMBL 65, miRbase18, RNA22v2	https://cm.jefferson.edu/datatoolsdownloads/rna22-full-sets-of-predic tions/

5.4 TARGETSCAN

TargetScan è un server web che contiene dati biologici di microRNA (miRNA) di molte specie: TargetScanHuman, TargetScanMouse, TargetScanFish, TargetScanFly e TargetScanWorm.

Le basi di dati citate forniscono previsioni per i mammiferi, Danio zebrato, insetti, nematodi, topo, Moscerino della frutta (Drosophila Melanogaster) e Caenorhabditis elegans. Rispetto ad altri strumenti di previsione, TargetScan fornisce classifiche accurate dei target di miRNA basate su una logica di contesti e di punteggi.

I dati sono forniti in un unico file in formato Tab Separated Values. Come prima operazione si è proceduto ad isolare la specie di interesse escludendo dal dataset i record che non avessero l'ID 10090 (corrispondente a Mus Musculus) come valore nella quinta colonna.



miR Family	Gene ID	Gene Symbol	Transcript ID	Species ID	UTR start	UTR end	MSA start	MSA end	Seed match	PCT
miR-15-5p/16-5p/195-5p/424-5p/497-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9544	466	472	729	736		
miR-326-3p/330-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10116	422	428	663	669	7mer-m8	NULL
miR-149	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9598	2010	2017	4067	4074	8mer	NULL
miR-149-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9606	2009	2016	4067	4074	8mer	NULL
miR-149	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9615	2159	2166	4067	4074	8mer	NULL
miR-493-3p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9606	578	584	903	909	7mer-m8	NULL
miR-493	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9913	460	466	903	909	7mer-m8	NULL
miR-383	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9544	647	654	990	1001	8mer	0.01
miR-96-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10116	1766	1773	3035	3042	8mer	0.00
miR-96-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10090	2020	2027	3690	3698	8mer	0.00
miR-493-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9544	182	189	345	352	8mer	NULL
miR-27-3p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10090	399	406	636	643	8mer	0.00
miR-124-3p.2	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10090	2547	2554	4960	4968	8mer	NULL
miR-491-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10090	1728	1735	3073	3080	8mer	NULL
miR-491-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9606	2260	2267	4536	4560	8mer	NULL
miR-129-3p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9606	493	499	797	806	7mer-m8	NULL
miR-326-3p/330-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	10090	424	430	663	669	7mer-m8	NULL
miR-326	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9913	273	279	663	669	7mer-m8	NULL
miR-15/16/195/424/497	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9598	462	468	729	736	7mer-m8	0.52
miR-15-5p/16-5p/195-5p/424-5p/497-5p/6838-5p	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9606	457	463	729	736	7mer	
miR-15/16/195/424-5p/497	ENSMUSG00000021252.6	0610007P14Rik	ENSMUST00000021676.6	9913	334	340	729	736	7mer	

Figura 10 TargetScan File Format

I campi di interesse per il database TargetScan sono riportati nella seguente tabella.

name	Fissata a 'TargetScan';
microRNA	Prima colonna. Il nome non contiene il prefisso della specie (ad esempio, 'mmu-') che deve essere quindi aggiunto utilizzando il riferimento alla specie (Species ID) per trovare una corrispondenza (case -insensitive) tra i nodi microRNA. Alcuni nomi contengono un '.' e vengono aggiunti in Neo4j perché sicuramente non presenti in miRBase;
Target	Seconda colonna, escludendo dal '.' in poi. Il codice Ensembl viene usato per trovare il nodo Target corrispondente. Se questo non esiste viene creato ottenendo le informazioni nell'ordine da NCBI, UniProt e ensembl.org;
score	Undicesima colonna (PCT);
source_microrna	Nome originale del microRNA presente in TargetScan;
source_target	Codice Ensembl (inizia con ENS) che include la parte dopo il '.'.

Dettagli Banca Dati Target Scan

Database	Descrizione	Download
TargetScan	Predicted (conserved) targets of conserved miRNA families. Includes positions on UTRs (without gaps) and UTR multiple sequence alignments (MSA; with gaps)	http://www.targetscan.org/mmu_71/mmu_71_data_download/Conserved_Family_Conserved_Targets_Info.txt.zip

5.5 MIRBASE

Il miRBase è un database di ricerca di sequenze miRNA. Ogni voce nel database rappresenta una porzione di trascrizione di miRNA, con informazioni relative alla posizione e alla sequenza. Il Registro miRBase fornisce un sistema centralizzato per l'assegnazione di nuovi nomi di geni microRNA.

I dati del miRBase vengono forniti per il download (file: miRNA.dat) in un formato testuale che specifica un blocco di informazioni per ogni record (microRNA) presente nella banca dati. Ogni blocco inizia con una riga contenente la stringa 'ID' seguita dall'identificativo unico (nome) del microRNA e termina con una riga contenente solo la stringa '//'. Le diverse specie vengono discriminate verificando che l'identificativo del microRNA abbia come prefisso quello della specie da isolare (ad esempio, 'mmu' per *Mus Musculus*).

```
CC   form from large-scale cloning studies [4].
XX
FH   Key             Location/Qualifiers
FH
FT   miRNA           7..28
FT                   /accession="MIMAT0000121"
FT                   /product="mmu-let-7g-5p"
FT                   /evidence=experimental
FT                   /experiment="cloned [1-4], Illumina [5-6]"
FT   miRNA           63..84
FT                   /accession="MIMAT0004519"
FT                   /product="mmu-let-7g-3p"
FT                   /evidence=experimental
FT                   /experiment="cloned [4], Illumina [5-6]"
XX
SQ   Sequence 88 BP; 21 A; 20 C; 26 G; 0 T; 21 other;
     ccaggcugag guaguaguuu guacaguuug agggucuaug auaccacccg guacaggaga        60
     uaacuguaca ggccacugcc uugccagg                                           88
//
ID   mmu-let-7i        standard; RNA; MMU; 85 BP.
XX
AC   MI0000138;
XX
DE   Mus musculus let-7i stem-loop
XX
RN   [1]
RX   PUBMED; 12007417.
RA   Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T;
RT   "Identification of tissue-specific microRNAs from mouse";
RL   Curr Biol. 12:735-739(2002).
```

Figura 11 miRNA file format

I campi di interesse per il database miRBase sono riportati nella seguente tabella.

name	riga che inizia con 'ID', il nome del microRNA è la prima stringa dopo il primo TAB (per il campo name vengono selezionati anche gli identificativi segnati come '/product=');
accession	riga che inizia con 'AC', il codice accession è la prima stringa dopo il primo TAB escludendo il ';' finale (per il campo accession vengono selezionati anche i codici segnati come '/accession=');
mirbase_link	viene semplicemente riportato il codice accession utile a costruire il link a miRBase.

Dettagli Banca Dati mirBase

Database	Descrizione	Download
miRBase	miRNA.dat (all published miRNA data in EMBL format)	ftp://mirbase.org/pub/mirbase/CURRENT/miRNA.dat.gz

5.6 MIRTARBASE

mirTarBase è una base di dati che contiene interazioni miRNA-Target, la base di dati è costituita da più di 470.000 elementi, ed è stata costruita manualmente catalogando articoli scientifici e risultati di studi ed esperimenti su miRNA. Un articolo rilevante che attesta la qualità di MirTarBase è: (miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database di Hsi-Yuan Huang Et Al., 2020)

La base di dati è adesso disponibile sul sito http://miRTarBase.cuhk.edu.cn/ in formato .xls (Microsoft Excel). Di seguito è riportata una figura estratta dall'articolo di Hsi-Yuan Huang Et Al che descrive la costruzione e le fonti dei dati inseriti all'interno di mirTarBase.

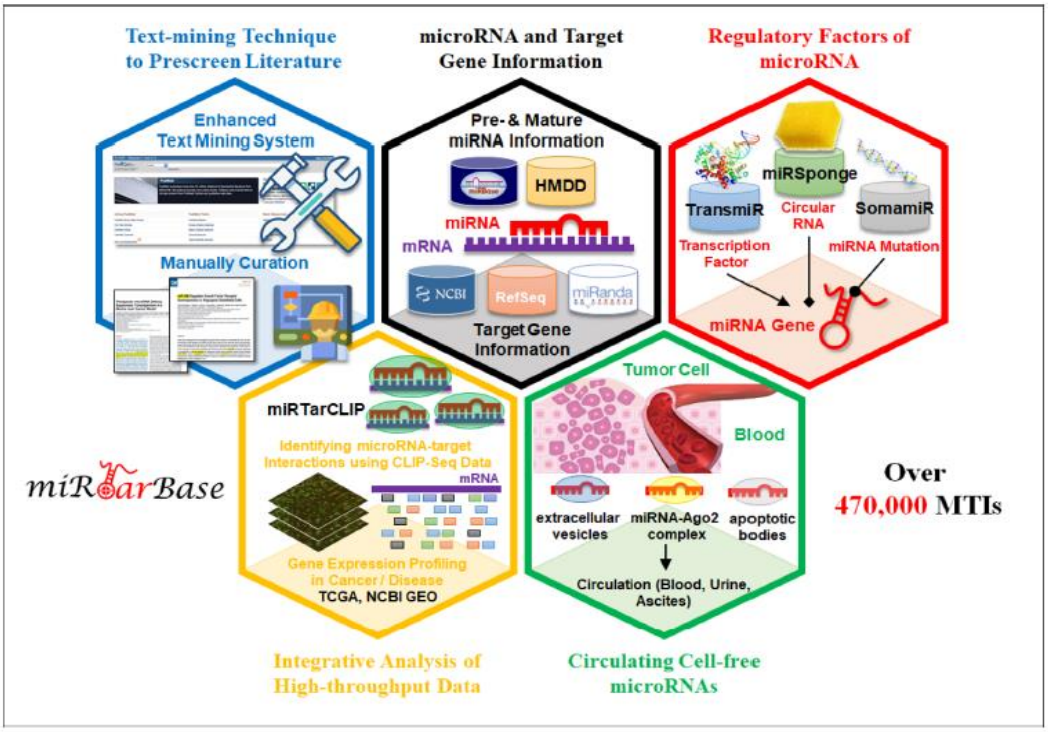


Figure 1. Highlighted improvements of miRTarBase 2020. As the most comprehensive resource on miRNA–target interactions, this update accumulates >470,000 manually confirmed MTIs supported with experimental evidence.

Table 1. List of the databases that are integrated by miRTarBase

Type	Database name
Gene and miRNA-specific databases	miRBase (7), NCBI Entrez gene (29), NCBI RefSeq (30)
SNP or mutation related databases	SomamiR (22,31)
miRNA–disease association Database	HMDD (9,10)
The regulation of miRNAs	TransMir (19), miRSponge (24)
miRNAs expression	CMEP (35), Gene Expression Omnibus (GEO) (32), The Cancer Genome Atlas (TCGA) (33,34)

Figura 12 struttura mirTarBase

Questa banca dati fornisce un file in formato Excel per ogni specie ('mmu_MTI.xls' per il Mus Musculus). Il file è stato convertito in formato Comma Separated Values per semplificare il processamento in Python. Nonostante il file sia specifico della specie Mus Musculus, sono presenti anche microRNA umani. Per discriminare la specie è quindi bastato selezionare solo i record il cui nome del microRNA cominciasse con il prefisso della specie in questione ('mmu_').

	A	B	C	D	E	F	G	H	I
1	miRTarBase ID	miRNA	Species (miRNA)	Target Gene	Target Gene	Species (Target	Experiments	Support Type	References (PMID)
2	MIRT053615	mmu-let-7a-5p	Mus musculus	ACVR1B	91	Mus musculus	Luciferase reporter as	Functional MTI	23152446
3	MIRT053616	mmu-let-7b-5p	Mus musculus	ACVR1B	91	Mus musculus	Luciferase reporter as	Functional MTI	23152446
4	MIRT437902	mmu-miR-155-5p	Mus musculus	CISH	1154	Mus musculus	Luciferase reporter as	Functional MTI	24778118
5	MIRT437843	hsa-miR-205-5p	Homo sapiens	ESRRG	2104	Mus musculus	Luciferase reporter as	Functional MTI	23589079
6	MIRT438043	mmu-let-7c-5p	Mus musculus	EZH2	2146	Mus musculus	Luciferase reporter as	Functional MTI	24365598
7	MIRT438710	mmu-let-7a-5p	Mus musculus	IL6	3569	Mus musculus	ELISA//Luciferase repo	Functional MTI	24001203
8	MIRT054373	mmu-miR-146a-5p	Mus musculus	IRAK1	3654	Mus musculus	qRT-PCR	Functional MTI (Weak)	22851573
9	MIRT054373	mmu-miR-146a-5p	Mus musculus	IRAK1	3654	Mus musculus	Luciferase reporter as	Functional MTI	24358114
10	MIRT054071	mmu-miR-128-3p	Mus musculus	PPARA	5465	Mus musculus	Luciferase reporter as	Functional MTI	22541023
11	MIRT437816	hsa-miR-558	Homo sapiens	PTGS2	5743	Mus musculus	Luciferase reporter as	Functional MTI	23611898
12	MIRT054881	mmu-miR-224-5p	Mus musculus	PTX3	5806	Mus musculus	Luciferase reporter as	Functional MTI	24470395
13	MIRT052983	mmu-miR-383-5p	Mus musculus	RBMS1	5937	Mus musculus	Luciferase reporter as	Functional MTI	22593182
14	MIRT437776	hsa-miR-1-3p	Mus musculus	TH	7054	Mus musculus	Luciferase reporter as	Functional MTI	25512392
15	MIRT052984	hsa-miR-92a-3p	Homo sapiens	TP53	7157	Mus musculus	Luciferase reporter as	Functional MTI	22451425
16	MIRT437354	mmu-miR-203-3p	Mus musculus	ZNF148	7707	Mus musculus	Luciferase reporter as	Functional MTI	22842794
17	MIRT054290	mmu-miR-106b-5p	Mus musculus	ULK1	8408	Mus musculus	Luciferase reporter as	Functional MTI	22781751
18	MIRT054289	mmu-miR-20a-5p	Mus musculus	ULK1	8408	Mus musculus	Luciferase reporter as	Functional MTI	22781751

Figura 13 miRTarBase file format

I campi di interesse sono stati selezionati in questo modo:

name	fissato a 'miRTarBase';
microRNA	seconda colonna. Non necessita di alterazioni. Le corrispondenze vengono trovate attraverso una ricerca case insensitive;
target	quinta colonna. Il nodo Target viene cercato utilizzando il GeneID. Se il nodo non esiste viene creato scaricando da NCBI le informazioni necessarie al popolamento degli attributi del nuovo nodo;
score	nona colonna (References PIMD);
source_microrna	contiene il nome del microRNA;
source_target	contiene il valore della quarta colonna (nome del gene).

Dettagli Banca Dati MirTarBase

Database	Descrizione	Download
miRTarBase	Mus musculus	http://mirtarbase.mbc.nctu.edu.tw/cache/download/6.1/mmu_MTI.xls

6 INTERFACCIA UTENTE

Il nuovo Diana Project rinnova le interfacce che possono essere utilizzate dagli utenti. In particolare è presente: una form che permette di interrogare il database e una seconda interfaccia che mostra i risultati delle interrogazioni in forma tabellare.

In modo similare all'articolo (Bioinformatics approach to predict target genes for dysregulated microRNAs in hepatocellular carcinoma: study on a chemically-induced HCC mouse model, 2015) è stata realizzata una tabella per visualizzare i risultati delle predizioni delle diverse banche dati. Di seguito è riportato un estratto dell'articolo che mostra la tabella originale.

L'utente può costruire query sul database partendo da un miRNA ed avere un risultato che comprende tutti i nodi target del miRNA di partenza, su ogni riga della tabella sono presenti 0 o più colori che indicano la presenza della relazione tra miRNA e Target sui diversi database.

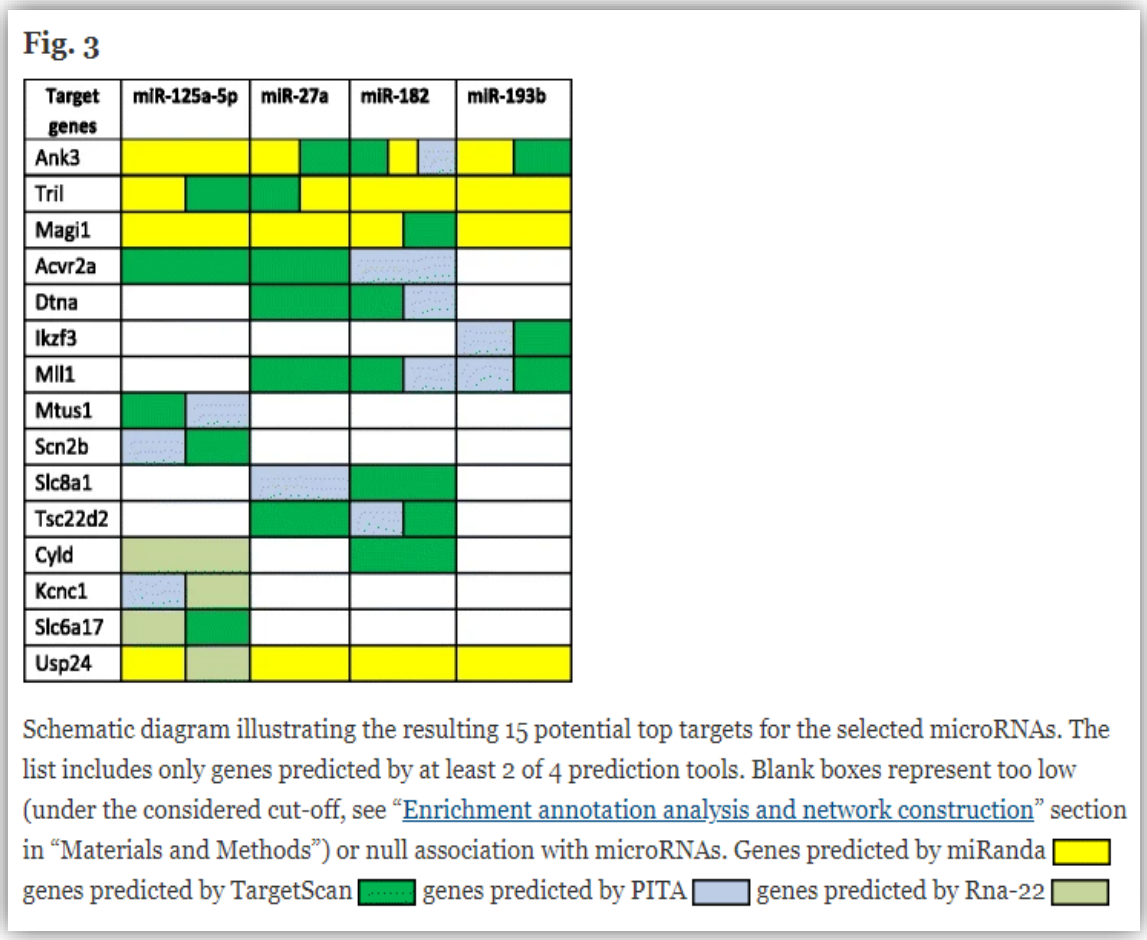


Fig. 3

Target genes	miR-125a-5p	miR-27a	miR-182	miR-193b
Ank3				
Tril				
Magi1				
Acvr2a				
Dtna				
Ikzf3				
Mll1				
Mtus1				
Scn2b				
Slc8a1				
Tsc22d2				
Cyld				
Kcnc1				
Slc6a17				
Usp24				

Schematic diagram illustrating the resulting 15 potential top targets for the selected microRNAs. The list includes only genes predicted by at least 2 of 4 prediction tools. Blank boxes represent too low (under the considered cut-off, see "Enrichment annotation analysis and network construction" section in "Materials and Methods") or null association with microRNAs. Genes predicted by miRanda genes predicted by TargetScan genes predicted by PITA genes predicted by Rna-22

Figura 14 Tabella di Predizione (Bioinformatics approach to predict target genes for dysregulated microRNAs in hepatocellular carcinoma: study on a chemically-induced HCC mouse model, 2015)

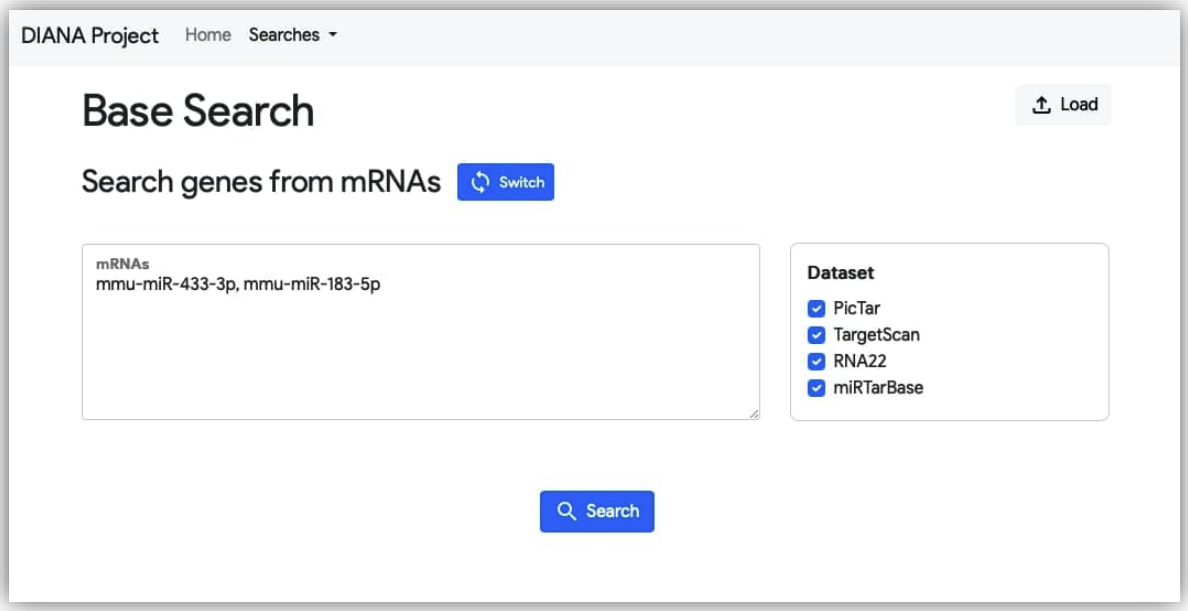


Figura 15 Form delle query da miRNA a Gene

La form realizzata con il framework Bootstrap permette agli utenti di fare interrogazioni inserendo il nome del miRNA inoltre permette di selezionare uno o più sorgenti di dati. L'utente è guidato nella costruzione delle query con un **controllo testuale implementato tramite RegEx** come mostrato nella figura seguente.

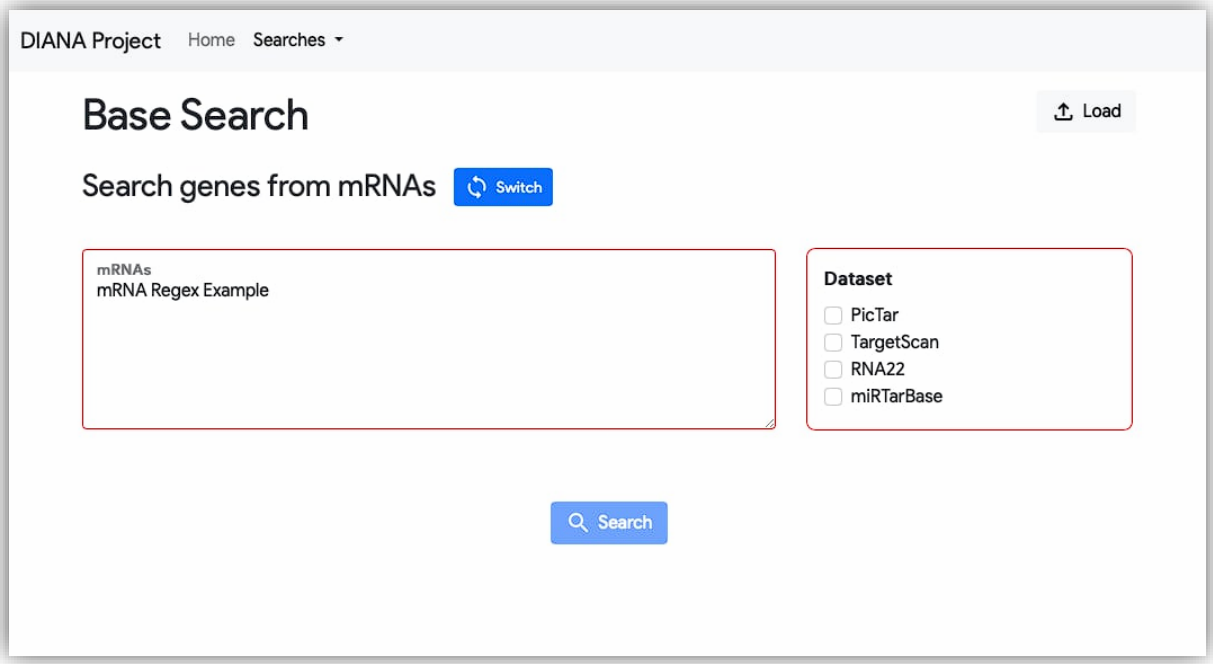


Figura 16 Esempio di controllo testuale RegEX

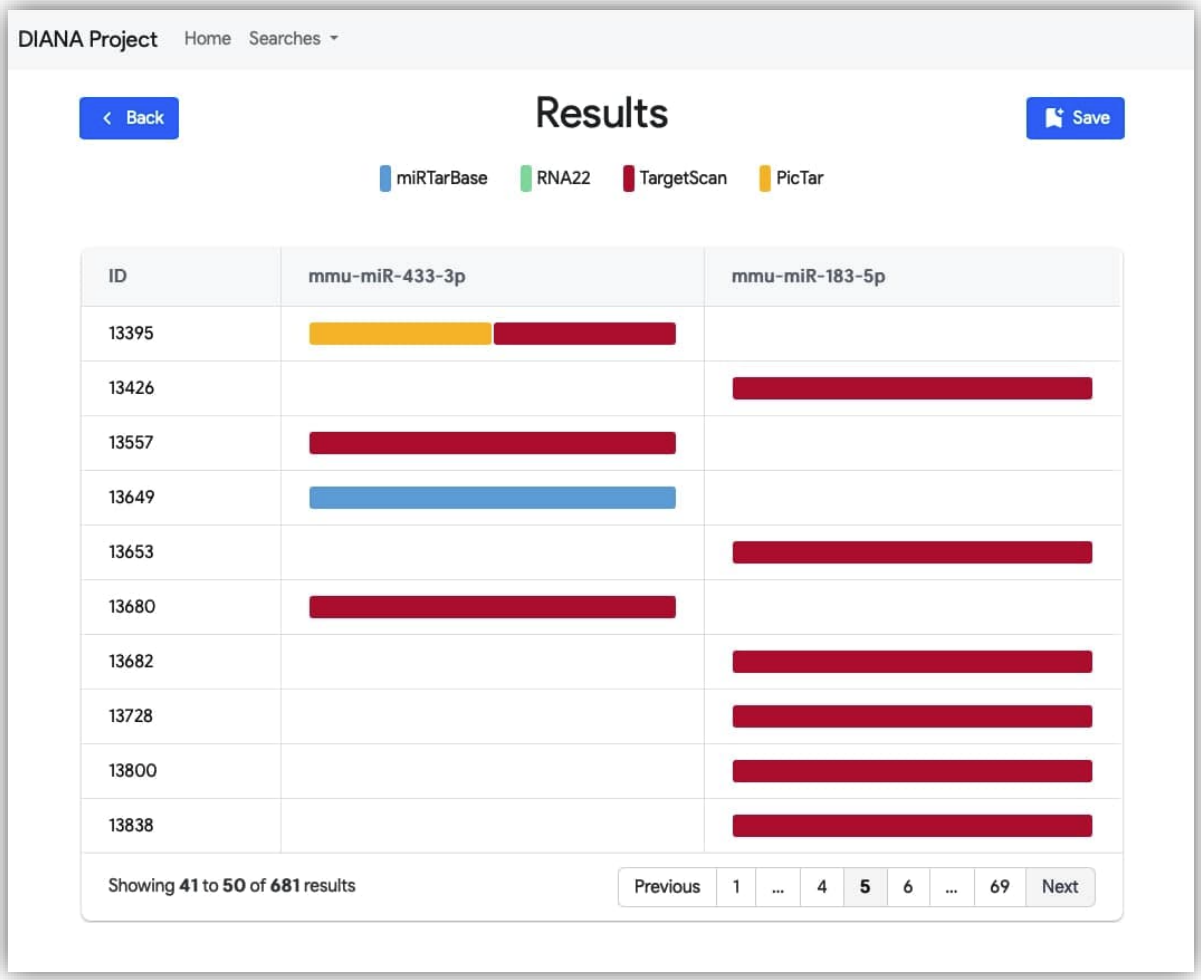


Figura 17 Nuova interfaccia tabellare per la visualizzazione dei risultati delle interrogazioni.

Il risultato mostrato nella figura precedente evidenzia la presenza della relazione su due diversi database, in questo caso riportato ad esempio la **relazione tra miRNA con id mmu-mir-433-3p ed il gene con id 13395 che è presente sia su TargetScan che su Pictar**.

Un'altra feature implementata permette agli utenti di interrogare il database a partire da un gene e da uno o più miRNA. In questo modo si permette all'utente di verificare la possibile relazione di tra uno o più miRNA ed un Gene target.

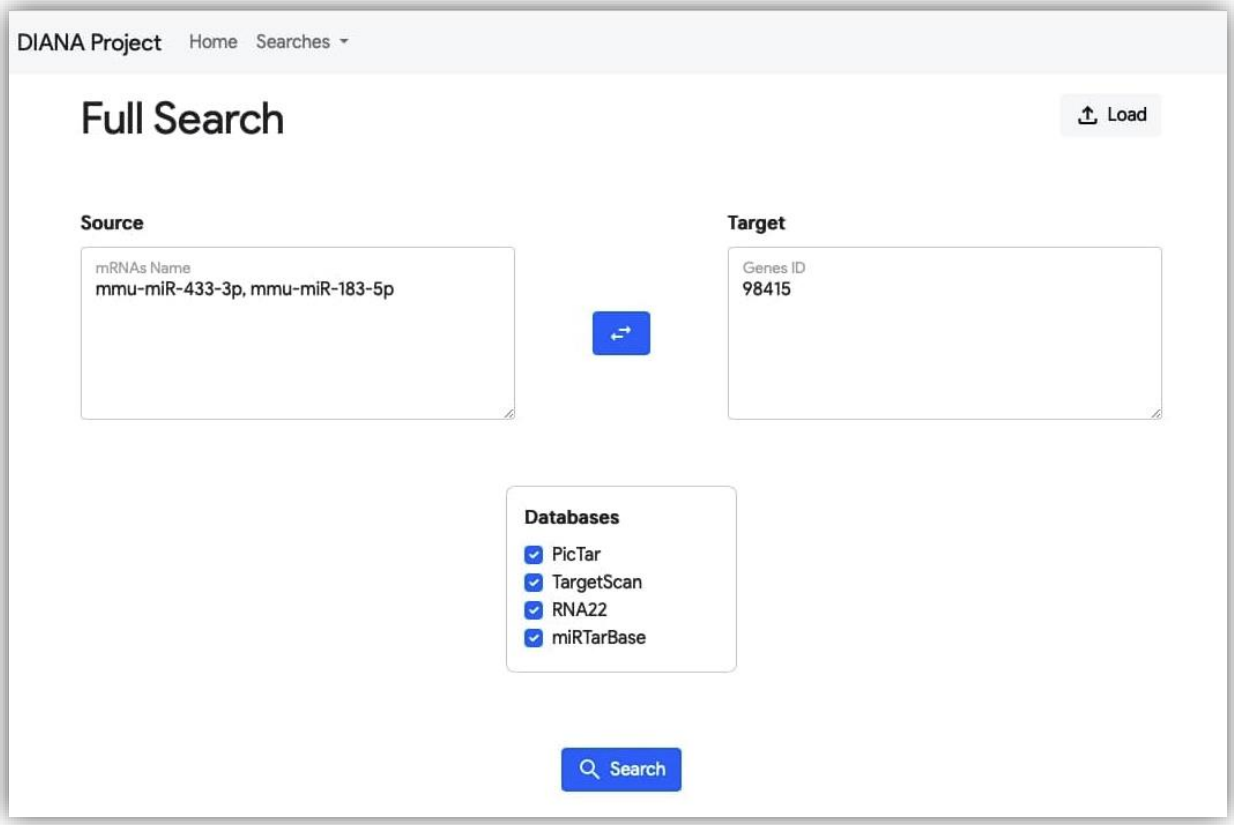


Figura 18 Form per interrogare il database che permette di inserire sia miRNA che Gene come parametri

Figura 19 Riusiltato della query con paramentri miRNA e GeneId

Infine, è stata implementata una feature che permette agli utenti di salvare le query effettuate in precedenza e caricarle in un secondo momento.

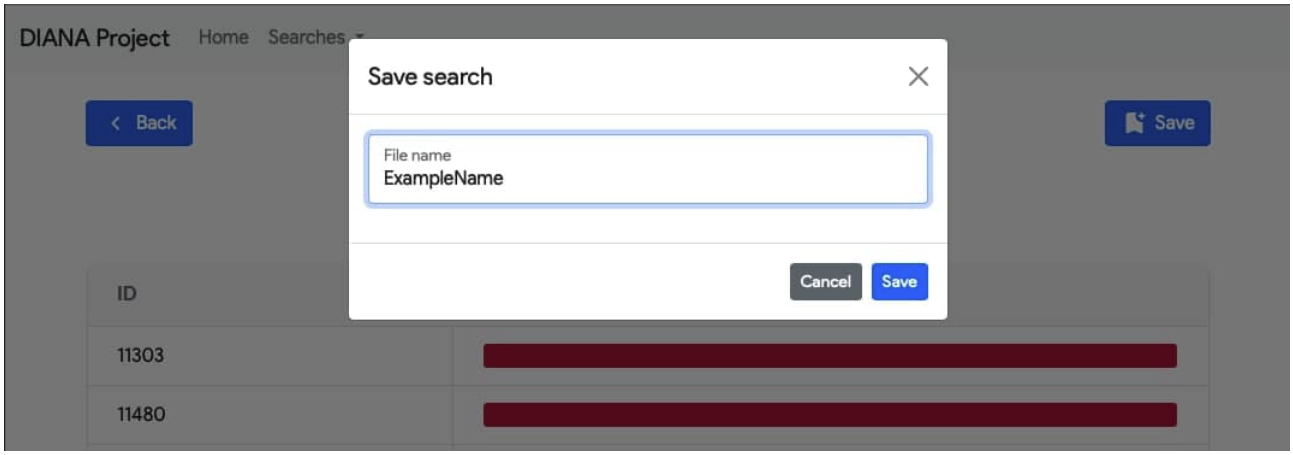


Figura 20 salvataggio query

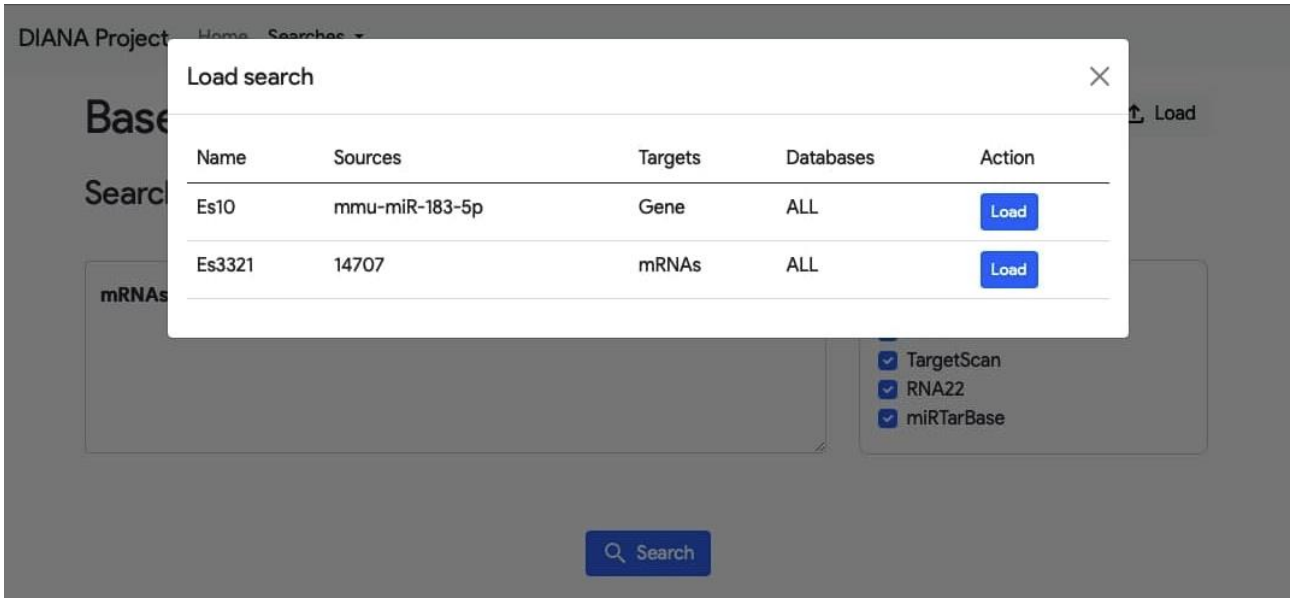


Figura 21 Caricamento query

7 CONCLUSIONI

Il lavoro effettuato durante questo progetto ha **migliorato l'architettura software** del Progetto Diana, questo miglioramento garantisce una migliore mantenibilità del software per i futuri sviluppi del progetto rendendo facilmente scalabili le implementazioni delle nuove feature. Fornisce una **componente View** che può essere estesa e modificata facilmente grazie all'utilizzo di **GridJS**, permettendo la gestione di tabelle dinamiche mediante **Javascript** ed il framework **Bootstrap** per creare agilmente nuove view e/o migliorare quelle già presenti senza la necessità di sviluppatori esperti.

Il micro **Framework Flask** permette di sviluppare facilmente e velocemente applicazioni web, sfruttando il linguaggio di programmazione Python, di comune utilizzo in ambito BioInformatico e già utilizzato nel Diana Project.

In futuro, dal punto di vista BioInformatico, è prevista l'integrazione con la base di dati Kegg, sviluppata dall'Università di Kyoto che contiene informazioni sui *pathway* metabolici della cellula, e focalizza l'attenzione sulle variazioni delle vie metaboliche tra diversi organismi viventi. (KEGG mapping tools for uncovering hidden features in biological data , 2021).

8 RISORSE ONLINE

1. **miRbase**: ftp://mirbase.org/pub/mirbase/CURRENT/miRNA.dat.gz
2. **UniProt**: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz
3. **Pictar**: https://genome.ucsc.edu/cgibin/hg
4. **mirTarBase**: *http://miRTarBase.cuhk.edu.cn/*
5. **TargetScan**: http://www.targetscan.org/mmu_71/mmu_71_data_download/Conserved_Family_Conserved_Targets_Info.txt.zip
6. **RNA22**: https://cm.jefferson.edu/datatoolsdownloads/rna22fullsetsofpredictions/
7. **GridJS**: https://gridjs.io/
8. **Flask:** https://palletsprojects.com/p/flask/
9. **RefSeq NCBI:** https://www.ncbi.nlm.nih.gov/books/NBK50679/#RefSeqFAQ.what_is_a_reference_sequence_r

9 INDICE DELLE FIGURE

10 BIBLIOGRAFIA

A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. **CHAD J. CREIGHTON, Et al. 2008.** s.l. : Cold Spring Harbor Laboratory Press., 2008.

A. di Marco, Et al. 2016. *A bioinformatics approach to predict the Influence of multiple conjoint mirnAs on cancer disease: the DIANA project.* L'Aquila : Università degli Studi dell'Aquila, 2016.

Are graph databases ready for bioinformatics? **Christian Theil Have, Lars Juhl Jensen. 2013.** 24, Copenhagen : BIOINFORMATICS Editorial, 2013, Vol. 29.

Bioinformatics approach to predict target genes for dysregulated microRNAs in hepatocellular carcinoma: study on a chemically-induced HCC mouse model. **Filippo Del Vecchio, Francesco Gallo, Antinisca Di Marco, Valentina Mastroiaco, Pasquale Caianiello, Francesca Zazzeroni, Edoardo Alesse, Alessandra Tessitore. 2015.** s.l. : BMC informatics, 2015.

KEGG mapping tools for uncovering hidden features in biological data . **Minoru Kanehisa, Et al. 2021.** s.l. : Institute for Chemical Research, Kyoto University, 2021.

KEGG: Kyoto Encyclopedia of Genes and Genomes. **Hiroyuki Ogata, Et al.** s.l. : Institute for Chemical Research, Kyoto University.

miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database di Hsi-Yuan Huang Et Al. **Hsi-Yuan Huang, Et al. 2020.** National Chiao Tung University : s.n., 2020.

Tucci Michele, Domenico Di Cesare,Federico Flaiano. 2016. *DIANA Project sorgenti di dati (documentazione Diana Project).* 2016.

UniProt: the Universal Protein knowledgebase. **Rolf Apweiler, Amos Bairoch,Winona C. Barker. 2004.** D115-D119, s.l. : Nucleic Acids Research, 2004, Vol. 32.